All You Need is Beyond a Good Init: Exploring Better Solution for Training Extremely Deep Convolutional Neural Networks with Orthonormality and Modulation

Di Xie
xiedi@hikvision.com

Jiang Xiong
xiongjiang@hikvision.com

Shiliang Pu
pushiliang@hikvision.com

Hikvision Research Institute
Hangzhou, China

1. Quasi-isometry inference with Batch Normalization

For batch normalization (BN) layer, its Jacobian, denoted as $\mathbf{J}$, is not only related with components of activations (d components in total), but also with samples in one mini-batch (size of m).

Let $x_j^{(k)}$ and $y_i^{(k)}$ be kth component of jth input sample and ith output sample respectively and given the independence between different components, $\frac{\partial y_i^{(k)}}{\partial x_j^{(k)}}$ is one of $m^2 d$ nonzero entries of $\mathbf{J}$. In fact, $\mathbf{J}$ is a tensor but we can express it as a blocked matrix:

$$\mathbf{J} = \begin{bmatrix} \mathbf{D}_{11} & \mathbf{D}_{12} & \cdots & \mathbf{D}_{1m} \\ \mathbf{D}_{21} & \mathbf{D}_{22} & \cdots & \mathbf{D}_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{D}_{m1} & \mathbf{D}_{m2} & \cdots & \mathbf{D}_{mm} \end{bmatrix} \tag{1}$$

where each $\mathbf{D}_{ij}$ is a $d \times d$ diagonal matrix:

$$\mathbf{D}_{ij} = \begin{bmatrix} \frac{\partial y_i^{(1)}}{\partial x_j^{(1)}} & & & \\ & \frac{\partial y_i^{(2)}}{\partial x_j^{(2)}} & & \\ & & \ddots & \\ & & & \frac{\partial y_i^{(d)}}{\partial x_j^{(d)}} \end{bmatrix} \tag{2}$$

Since BN is a component-wise rather than sample-wise transformation, we prefer to analyse a variant of Eq. 1 instead of $\mathbf{D}_{ij}$. Note that by elementary matrix transformation, the m^2 $d \times d$ matrices can be converted into d $m \times m$ matrices:

$$\mathbf{J} = \begin{bmatrix} \mathbf{J}_{11} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{J}_{22} & \cdots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \mathbf{J}_{dd} \end{bmatrix} \tag{3}$$

and the entries of each $\mathbf{J}_{kk}$ is

$$\frac{\partial y_j}{\partial x_i} = \rho \left[\Delta(i=j) - \frac{1 + \hat{x}_i \hat{x}_j}{m} \right] \tag{4}$$

The notations of ρ, $\Delta(\cdot)$ and $\hat{x}_k$ have been explained in our main paper and here we omit the component index k for clarity. Base on the observation of Eq. 4, we separate the numerator of latter part and denote it as $U_{ij} = 1 + \hat{x}_i \hat{x}_j$.

Let $\hat{\mathbf{x}} = (\hat{x}_1, \hat{x}_2, ..., \hat{x}_m)^T$, $\mathbf{e} = (1, 1, ..1)^T$, we have

$$\mathbf{U} = \mathbf{e}\mathbf{e}^T + \hat{\mathbf{x}}\hat{\mathbf{x}}^T \tag{5}$$

and

$$\mathbf{J}_{kk} = \rho(\mathbf{I} - \frac{1}{m}\mathbf{U}) \tag{6}$$

Recall that for any column vector $\mathbf{v}$, $\texttt{rank}(\mathbf{v}\mathbf{v}^T) = 1$. According to the subadditivity of matrix rank [1], it implies that

$$\begin{aligned} \texttt{rank}(\mathbf{U}) = \texttt{rank}(\mathbf{e}\mathbf{e}^T + \hat{\mathbf{x}}\hat{\mathbf{x}}^T) \leq \\ \texttt{rank}(\mathbf{e}\mathbf{e}^T) + \texttt{rank}(\hat{\mathbf{x}}\hat{\mathbf{x}}^T) = 2 \end{aligned} \tag{7}$$

Eq. 7 tells us that $\mathbf{U}$ actually only has two nonzero eigenvalues, say λ_1 and λ_2, and we can formulate $\mathbf{U}$ as follow:

$$\mathbf{U} = \mathbf{P}^T \begin{bmatrix} \lambda_1 & & & & \\ & \lambda_2 & & & \\ & & 0 & & \\ & & & \ddots & \\ & & & & 0 \end{bmatrix} \mathbf{P} \tag{8}$$

combined with Eq. 6, finally we get the equation of $\mathbf{J}_{kk}$ from the eigenvalue decomposition view, which is

$$\mathbf{J} = \mathbf{P}^T \rho \begin{bmatrix} 1 - \frac{\lambda_1}{m} & & & & \\ & 1 - \frac{\lambda_2}{m} & & & \\ & & 1 & & \\ & & & \ddots & \\ & & & & 1 \end{bmatrix} \mathbf{P} \tag{9}$$

To show that $\mathbf{J}_{kk}$ probably is not full rank, we formulate the relationship between $\mathbf{U}^2$ and $\mathbf{U}$

$$\begin{aligned}\mathbf{U}^2 = (\mathbf{e}\mathbf{e}^T + \hat{\mathbf{x}}\hat{\mathbf{x}}^T)(\mathbf{e}\mathbf{e}^T + \hat{\mathbf{x}}\hat{\mathbf{x}}^T) = \mathbf{e}\mathbf{e}^T\mathbf{e}\mathbf{e}^T + \mathbf{e}\mathbf{e}^T\hat{\mathbf{x}}\hat{\mathbf{x}}^T \\ + \hat{\mathbf{x}}\hat{\mathbf{x}}^T\mathbf{e}\mathbf{e}^T + \hat{\mathbf{x}}\hat{\mathbf{x}}^T\hat{\mathbf{x}}\hat{\mathbf{x}}^T = m\mathbf{e}\mathbf{e}^T + (\sum_{i=1}^{m}\hat{x}_i)\mathbf{e}\hat{\mathbf{x}}^T \\ + (\sum_{i=1}^{m}\hat{x}_i)\hat{\mathbf{x}}\mathbf{e}^T + (\sum_{i=1}^{m}\hat{x}_i^2)\hat{\mathbf{x}}\hat{\mathbf{x}}^T \\ = m\mathbf{U} + (\sum_{i=1}^{m}\hat{x}_i)\mathbf{e}\hat{\mathbf{x}}^T + (\sum_{i=1}^{m}\hat{x}_i)\hat{\mathbf{x}}\mathbf{e}^T + (\sum_{i=1}^{m}\hat{x}_i^2 - m)\hat{\mathbf{x}}\hat{\mathbf{x}}^T\end{aligned} \tag{10}$$

Note that $\hat{x}_i \sim N(0,1)$, so we can regard the one-order and second-order accumulated items in Eq. 10 as approximately equaling the corresponding one-order and second-order statistical moments for relatively large mini-batch, from which we get $\mathbf{U}^2 \approx m\mathbf{U}$.

The relationship implies that $\lambda_1^2 \approx m\lambda_1$ and $\lambda_2^2 \approx m\lambda_2$. Since λ_1 and λ_2 cannot be zeros, it concludes that $\lambda_1 \approx \lambda_2 \approx m$ therefor $1 - \frac{\lambda_1}{m} \approx 0$ and $1 - \frac{\lambda_2}{m} \approx 0$ if batch size is sufficient in a statistical sense.

References

[1] S. Banerjee and A. Roy. Linear algebra and matrix analysis for statistics. *Crc Press*, 2014.